W362–W365 *Nucleic Acids Research, 2006, Vol. 34, Web Server issue*
doi:10.1093/nar/gkl124

ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins

Edouard de Castro[1,*], Christian J. A. Sigrist[1], Alexandre Gattiker[3], Virginie Bulliard[1], Petra S. Langendijk-Genevaux[1], Elisabeth Gasteiger[1], Amos Bairoch[1,2] and Nicolas Hulo[1]

[1]Swiss Institute of Bioinformatics (SIB), 1 rue Michel Servet, CH-1211 Geneva 4, Switzerland, [2]Structural Biology and Bioinformatics Department, University of Geneva, 1 rue Michel Servet, CH-1211 Geneva 4, Switzerland and [3]National Institute of Technology and Evaluation, 2-49-10 Nishihara, Shibuya-ku, Tokyo 151-0066, Japan

Received February 14, 2006; Revised March 1, 2006; Accepted March 13, 2006

ABSTRACT

ScanProsite—http://www.expasy.org/tools/scanprosite/—is a new and improved version of the web-based tool for detecting PROSITE signature matches in protein sequences. For a number of PROSITE profiles, the tool now makes use of ProRules—context-dependent annotation templates—to detect functional and structural intra-domain residues. The detection of those features enhances the power of function prediction based on profiles. Both user-defined sequences and sequences from the UniProt Knowledgebase can be matched against custom patterns, or against PROSITE signatures. To improve response times, matches of sequences from UniProtKB against PROSITE signatures are now retrieved from a pre-computed match database. Several output modes are available including simple text views and a rich mode providing an interactive match and feature viewer with a graphical representation of results.

INTRODUCTION

To predict protein function, assign family identity or detect remote homologues, searches against signature databases, also known as secondary databases (1), are essential. ScanProsite provides a web interface to identify protein matches against signatures from the PROSITE database (2).

The PROSITE database consists of a large collection of biologically meaningful signatures that are described as patterns (regular expressions), used for short motif detection, or generalized profiles (weight matrices) for sensitive detection of larger domains. All signatures are built from manually derived alignments and are provided with extensive manually curated documentation (taxonomic occurrence, function, etc.). PROSITE signatures form a high quality collection and are closely tied to the UniProtKB/Swiss-Prot (3,4) annotation process.

In addition, each PROSITE profile is associated with a manually curated annotation template called ProRule (5). These rules are used internally, by the Swiss-Prot group, to automate PROSITE domain annotation of UniProtKB/Swiss-Prot entries. A number of ProRules define biologically meaningful information about specific residues within their associated domain. This positional information (derived from the mapping of significant residues to the profile) is provided in the form of contextual feature annotation blocks: certain conditions—a specific sequence, the presence of other features—must be fulfilled for the annotation block to be applied, hence for the feature to be predicted.

Consequently, those ProRules add pattern-like discriminativity with motif-specific information to their associated profile, allowing the detection of intra-domain features, such as active sites, binding sites or disulfide bridges. Combining the sensitivity of profiles with the specificity of motif detection enhances the accuracy of signature based functional predictions.

The ScanProsite rich result viewer detects those intra-domain features by evaluating—on the fly—associated ProRules on matched profiles and integrates them with the match results.

PROSITE is an InterPro (6) database member. The InterProScan tool (http://www.ebi.ac.uk/InterProScan/) can scan sequences against PROSITE signatures ('ProfileScan' application), but does not provide intra-domain feature detection, 'rich' graphical output and extensive scanning options including scans against custom user patterns.

*To whom correspondence should be addressed. Tel: +41 22 379 5050; Fax: +41 22 379 5858; Email: ecastro@isb-sib.ch

DESIGN AND IMPLEMENTATION

ScanProsite was implemented in Perl, and is served through an Apache web server running on a UNIX operating system. Care was taken to ensure that all generated pages are fully standards-compliant (valid HTML 4.01 transitional). The pre-computed match database and the ProRule database are stored in a PostgreSQL database. The entire implementation is based on open source tools.

Program output can either be displayed in the web browser (interactive mode), or sent by email (batch mode). The rich view output mode uses standard DHTML (HTML, CSS, JavaScript), and no additional plugins are required.

Input form

The job submission starting page is located at http://www.expasy.org/tools/scanprosite/ on the ExPASy web server (7).

Here the user has to enter sequence and/or signature data, choose the output format, and specify various scan behavior options before launching the analysis job. The form usage is described at http://www.expasy.org/tools/scanprosite/scanprosite-doc.html.

User sequences and/or UniProtKB sequences/databases can be scanned against user patterns and/or PROSITE signatures/database. Multiple sequences and/or signatures can be submitted at once (maximum: 8 for scans against databases, otherwise 16). In this case, any specified signature will be searched against any specified sequence (logical OR).

For scans against protein databases, the search space can be reduced to specific taxa, and/or to entries containing a specific term in their description (UniProtKB DE line).

Protein databases can be randomized (reversed or shuffled) to evaluate user pattern significance (8). Pattern matching mode can be altered (see aforementioned online documentation page under 'pattern matching mode').

Users can choose to retrieve full sequences (in fasta format) of matched proteins together with signature match results.

For scans against the whole PROSITE signature database, 'unspecific' signatures with a high probability of occurrence (see PROSITE user manual at http://www.expasy.org/prosite/prosuser.html) can be excluded (this option is activated by default). Scans can also be restricted to patterns.

Users can also choose to include low level profile matches (with score smaller than the high confidence - level 0 - cut-off) in the output (this option is not activated by default).

To save bandwidth and computing power, there is a limit on the number of distinct matching sequences and total matches that can be displayed. Users can select the maximum of matching sequences (by default, not more than 5000 matches in 1000 sequences will be displayed). A maximum of 10 000 or 100 000 sequences (50 000 or 500 000 signature matches) can only be requested in the non-interactive 'batch mode' where results are sent to the user via email.

The non-interactive 'batch mode' is used automatically when an email address is specified.

Instructions for simple programmatic access (text/plain output) are available on demand (email: prosite@expasy.org).

Sequence analysis

Once the job has been submitted, the data are posted to an application that will make use of pre-computed matches whenever possible. Results are retrieved from an internal relational database that stores matches of all PROSITE signatures (except the 'unspecific' signatures with a high probability of occurrence) against the UniProt Knowledgebase, including additional splice variants, and PDB (9). For analyses where no pre-computed results are available (e.g. on user sequences, user patterns or 'unspecific' signatures) a real-time sequence analysis is performed using the ps_scan program (10). Users requiring high-throughput sequence analysis or who wish to use custom databases can download ps_scan stand-alone tool at ftp://ftp.expasy.org/databases/prosite/tools/ps_scan/.

Available results are pooled and saved (for 12 h) on the server, for later use by the rich viewer.

The detection of intra-domain features is performed on the fly by the rich viewer over the matching domains (and is therefore only available in the rich view output mode, see below). Profile match regions are evaluated for fulfillment of feature detection conditions specified in the associated ProRules. The features themselves are represented in the ProRule as UniProtKB/Swiss-Prot annotation blocks that are applied when detection conditions are fulfilled. The feature annotation blocks and their detection conditions are retrieved from an internal database storing a relational representation of the ProRules.

Detection conditions can be: specific amino acids inside the domain (regular expressions that can be grouped by logical operators), groups of conditions in which all conditions must be fulfilled (e.g. catalytic triad of trypsin protease).

To date, intra-domain feature detection is performed for 136 of the 595 PROSITE profiles (PROSITE release 19.20, of 07-Feb-2006). Predicted features include (Table 1): binding sites for chemical groups (such as Heme, ATP), glycosylation site, disulfide bridges, DNA binding sites, metal ion binding sites, post-translational modification sites (such as phosphotyrosine, phosphoserine) etc. The system can be easily extended to use other UniProtKB/Swiss-Prot feature types.

ScanProsite output

In interactive mode, once the analysis is completed, the results will be directly displayed in the selected output view mode inside user's web browser.

Table 1. UniProtKB/Swiss-Prot features that can be predicted through ScanProsite (PROSITE release 19.20, of 07-Feb-2006)

UniProtKB/Swiss-Prot feature key name[a]	No. of profiles	Example (profile AC/ID)
ACT_SITE	26	PS50240/TRYPSIN_DOM
BINDING	8	PS51007/CYTC
CA_BIND	1	PS50222/EF_HAND_2
CARBOHYD	2	PS50015/SAP_B
DISULFID	45	PS50948/PAN
DNA_BIND	17	PS51063/HTH_CRP_2
METAL	24	PS50873/PEROXIDASE_4
MOD_RES	20	PS51149/GLY_RADICAL_2
NP_BIND	9	PS50936/ENGC_GTPASE
SITE	2	PS51062/RUNT
ZN_FING	6	PS50115/ARFGAP

[a]See UniProtKB/Swiss-Prot user manual at http://www.expasy.org/sprot/userman.html#FTID.

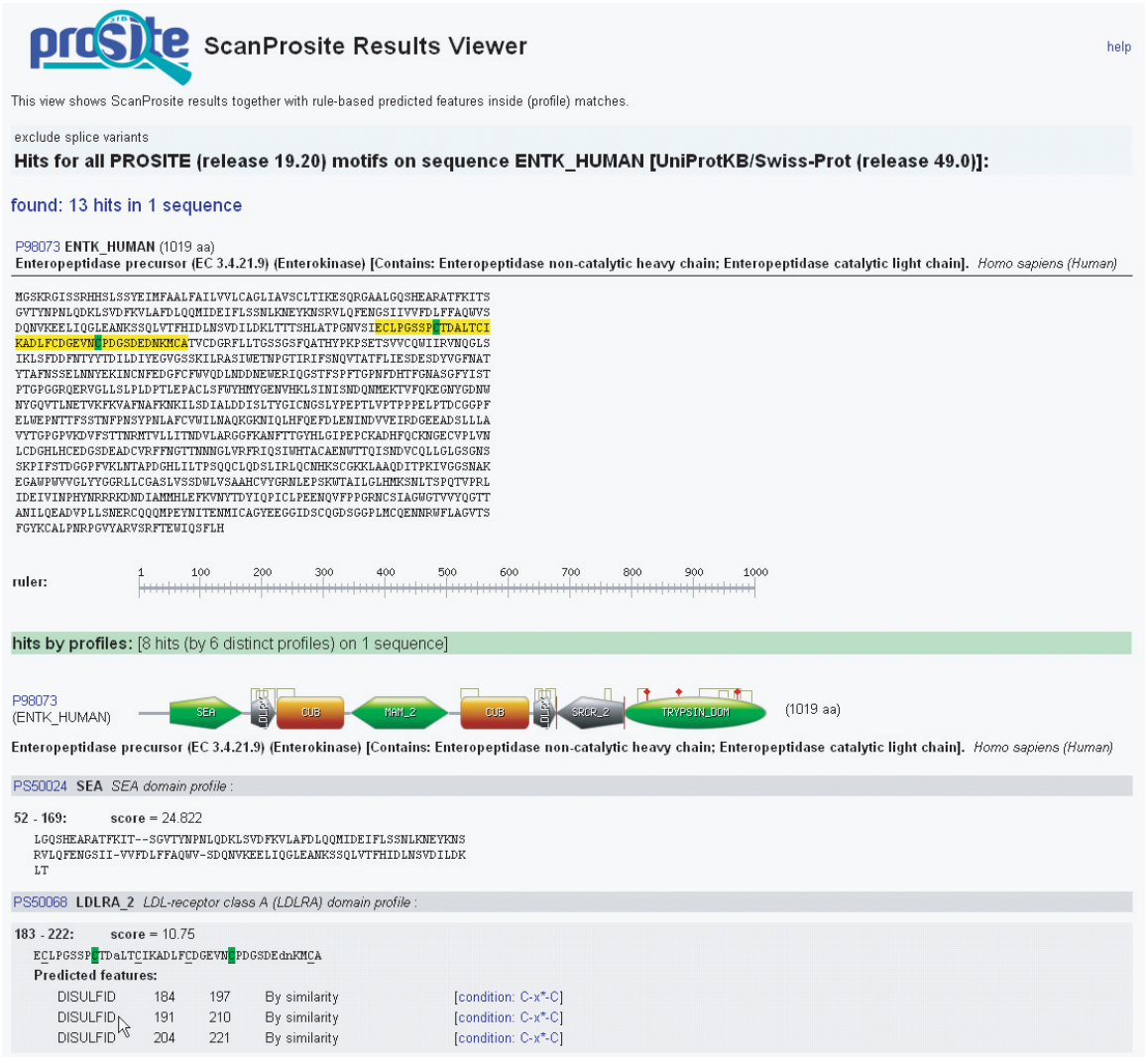

Figure 1. ScanProsite result page (rich view mode).

In batch mode, results in simple text format are emailed to the user-specified email address, together with a link to visualize the results, stored on the server, through the rich viewer. In rich view and simple html view modes, matching UniProt Knowledgebase protein entries are shown with a link to their ExPASy NiceProt view, their description and organism. Moreover, for entries that have associated PDB structures, there are links to interactive 3D views highlighting the match region over the 3D structure.

Text modes provide no links, but are optimal for copy/pasting. See http://www.expasy.org/tools/scanprosite/scanprosite-doc.html#output for details.

The rich view can be accessed, as a link, from any other views.

The rich view output mode provides an interactive match viewer and intra-domain feature predictor/viewer in both text and graphical forms. See http://www.expasy.org/tools/scanprosite/scanview-doc.html for details.

The graphical view is a symbolic representation of matches, inspired by SMART (11) and predicted features in a high quality image (Portable Network Graphics format) that can be used in presentations or papers (Figure 1).

The view is generated by a specific web tool that uses saved results allowing delayed results examination for up to 12 h after the initial scan (useful in batch mode).

In addition to the match results, intra-domain feature evaluations (if any) are detailed. For each evaluated feature, its UniProtKB/Swiss-Prot feature key, boundaries, description, the detection conditions and the condition group (if any), are displayed. Features with fulfilled conditions are shown under 'Predicted features' and integrated into the graphical representation. Features with unfulfilled conditions are shown under 'Absent feature'.

The rich view also provides interactive match and feature highlighting, on results of a single sequence scan, with most web browsers e.g. Mozilla, FireFox, Netscape, Opera, provided JavaScript is enabled (with Internet Explorer this JavaScript functionality is too slow and was disabled). Moving the cursor over a feature information line will highlight its position (green for predicted features, gray for absent features)

in both the match sequence and the protein sequence. Moving the cursor over a match in the graphical view or the text view will highlight in yellow its position in the protein sequence.

CONCLUSIONS

We have described the current state of the ScanProsite tool. The new implementation—online since summer 2004—brings the use of pre-computed matches whenever possible, the detection of intra-domain biological features, a new graphical result representation and an interactive result viewer.

PROSITE, through ScanProsite, provides broad intra-domain feature prediction via a flexible context-dependent annotation transfer system. Associating domain detection with an automated annotation system can significantly increase functional predictive power of profiles.

ACKNOWLEDGEMENTS

The authors would like to thank Eric Jain for careful proof-reading. This work was supported by grant no. 3152A0-103922/1 from the Swiss National Science Foundation and by the Swiss Federal Government through the Federal Office of Education and Science. Funding to pay the Open Access publication charges for this article was provided by Swiss Institute of Bioinformatics (Swiss-Prot group), Switzerland.

Conflict of interest statement. None declared.

REFERENCES

1. Attwood,T.K. and Parry-Smith,D.J. (1999) *Introduction to Bioinformatics*. Addison Wesley Longman.
2. Hulo,N., Bairoch,A., Bulliard,V., Cerutti,L., De Castro,E., Langendijk-Genevaux,P.S., Pagni,M. and Sigrist,C.J.A. (2006) The PROSITE database. *Nucleic Acids Res.*, **34**, D227–D230.
3. Bairoch,A., Boeckmann,B., Ferro,S. and Gasteiger,E. (2004) Swiss-Prot: juggling between evolution and stability. *Brief Bioinform.*, **5**, 39–55.
4. Wu,C.H., Apweiler,R., Bairoch,A., Natale,D.A., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R. *et al*. (2006) The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res.*, **34**, D187–D191.
5. Sigrist,C.J.A., De Castro,E., Langendijk-Genevaux,P.S., Le Saux,V., Bairoch,A. and Hulo,N. (2005) ProRule: a new database containing functional and structural information on PROSITE profiles. *Bioinformatics*, **21**, 4060–4066.
6. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Binns,D., Bradley,P., Bork,P., Bucher,P., Cerutti,L. *et al*. (2005) InterPro, progress and status in 2005. *Nucleic Acids Res.*, **33**, D201–D205.
7. Gasteiger,E., Gattiker,A., Hoogland,C., Ivanyi,I., Appel,R.D. and Bairoch,A. (2003) ExPASy: the proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res.*, **31**, 3784–3788.
8. Hulo,N., Sigrist,C.J.A., Le Saux,V., Langendijk-Genevaux,P.S., Bordoli,L., Gattiker,A., De Castro,E., Bucher,P. and Bairoch,A. (2004) Recent improvements to the PROSITE database. *Nucleic Acids Res.*, **32**, D134–D137.
9. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
10. Gattiker,A., Gasteiger,E. and Bairoch,A. (2002) ScanProsite: a reference implementation of a PROSITE scanning tool. *Appl. Bioinformatics*, **1**, 107–108.
11. Letunic,I., Copley,R.R., Pils,B., Pinkert,S., Schultz,J. and Bork,P. (2006) SMART 5: domains in the context of genomes and networks. *Nucleic Acids Res.*, **34**, D257–D260.